ΚΥΚΛΟΣ ΣΕΜΙΝΑΡΙΩΝ ΣΤΑΤΙΣΤΙΚΗΣ ΟΚΤΩΒΡΙΟΣ 2015

Απόστολος Μπουρνέτας
Τμήμα Μαθηματικών
Πανεπιστήμιο Αθηνών

Το πρόβλημα multi-armed-bandit και επεκτάσεις

ΤΕΤΑΡΤΗ 21/10/2015
15:00 – 17:00

ΑΙΘΟΥΣΑ 607, 6ος ΟΡΟΦΟΣ,
ΚΤΙΡΙΟ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ
(ΕΥΕΛΠΙΔΩΝ & ΛΕΥΚΑΔΟΣ)

ΠΕΡΙΛΗΨΗ

Θεωρούμε ένα πρόβλημα προσαρμοστικής δειγματοληψίας από ένα πεπερασμένο αριθμό ανεξάρτητων στατιστικών πληθυσμών, κάτω από ελλιπή πληροφόρηση σχετικά με τις υποκείμενες κατανομές πιθανότητας. Σε κάθε περίοδο επιλέγεται ένας πληθυσμός για τη λήψη μιας παρατήρησης. Το αποτέλεσμα της παρατήρησης θεωρείται ως αμοιβή και ο αντικειμενικός σκοπός είναι να αναπτυχθεί μια προσαρμοστική πολιτική δειγματοληψίας η οποία σε κάθε περίοδο επιλέγει από ποιον πληθυσμό θα πάρει την παρατήρηση, με βάση την προηγούμενη ιστορία επιλογών και παρατηρήσεων, έτσι ώστε να μεγιστοποιηθεί (υπό κατάλληλη ασυμπτωτική έννοια) η αναμενόμενη συνολική αμοιβή.

Σε Μπεϋζιανό πλαίσιο το πρόβλημα μπορεί να θεωρηθεί ως ειδική περίπτωση μιας γενικής κατηγορίας προβλημάτων προσαρμοστικής βελτιστοποίησης, γνωστής ως multi-armed-bandit problems. Για τα προβλήματα αυτά με κριτήριο την αναμενόμενη αποπληθωρισμένη αμοιβή σε άπειρο ορίζοντα έχει αποδειχθεί η βελιστότητα μιας κλάσης πολιτικών που βασίζονται σε δείκτες και επιτρέπουν τη διάσπαση του πολυδιάστατου προβλήματος σε ένα σύνολο μονοδιάστατων προβλημάτων βελτιστοποίησης. Η διάσπαση αυτή δεν είναι γενικά δυνατή όταν μετακινηθούμε σε μη Μπεϋζιανά μοντέλα ή/και σε κριτήρια διαφορετικά της αποπληθωρισμένης αμοιβής σε άπειρο ορίζοντα.

Στην ομιλία θα θεωρήσουμε ένα μη Μπεϋζιανό μοντέλο με κριτήριο τη μεγιστοποίηση του μέσου κόστους σε άπειρο ορίζοντα. Θα δείξουμε ότι το πρόβλημα μεγιστοποίησης είναι ισοδύναμο με την ελαχιστοποίηση κατάλληλης συνάρτησης απώλειας, και θα αποδείξουμε την ύπαρξη προσαρμοστικών πολιτικών που έχουν δομή δείκτη και εγγυώνται τον ελάχιστο ασυμπτωτικό ρυθμό αύξησης της απώλειας ή ισοδύναμα το μέγιστο ρυθμό σύγκλισης της μέσης αμοιβής σε αυτή που μπορεί να επιτευχθεί κάτω από πλήρη πληροφόρηση.

Επίσης θα θεωρήσουμε δύο επεκτάσεις του προβλήματος: (α) όταν υπάρχουν περιορισμοί στις συχνότητες δειγματοληψίας λόγω κόστους και (β) όταν οι αμοιβές παρουσιάζουν χρονική μη στασιμότητα υπό κατάλληλα ορισμένη έννοια.

Η ομιλία βασίζεται σε εργασίες από κοινού με τους Μιχάλη Κατεχάκη και Οδυσσέα Καναβέτα.

ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

ΣΧΟΛΗ ΕΠΙΣΤΗΜΩΝ & ΤΕΧΝΟΛΟΓΙΑΣ ΤΗΣ ΠΛΗΡΟΦΟΡΙΑΣ
SCHOOL OF INFORMATION SCIENCES & TECHNOLOGY

ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ
DEPARTMENT OF STATISTICS

AUEB STATISTICS SEMINAR SERIES OCTOBER 2015

Apostolos Burnetas
Department of Mathematics
University of Athens

Optimal Adaptive Policies in the Multi-Armed-Bandit Problem and Extensions

Wednesday 21/10/2015
15:00 – 17:00

ROOM 607, 6th FLOOR,
POSTGRADUATE STUDIES BUILDING
(EVELPIDON & LEFKADOS)

ABSTRACT

Consider the problem of adaptive sampling from a finite number of independent statistical populations under incomplete information on the underlying probability distributions. In every period one of the populations is selected to receive one observation from. The outcome of the observation is considered a reward. The objective is to develop an adaptive sampling policy which selects the population to sample from in each period, based on the previous history of selections and observations, so that the expected total reward is maximized under an appropriate asymptotic sense.

In a Bayesian framework the problem can be viewed as a special case of a general class of adaptive optimization problems, referred to in the literature as multi-armed-bandit problems. Under the criterion of maximizing the expected total discounted reward in infinite horizon it has been proved that the optimal sampling policy has a simple index structure. Specifically for each population there exists an index, which is a function of the previous information about this population, such that in every period it is optimal to sample from the population with the highest index value. This is an important result, because it allows decomposing the multi-dimensional dynamic optimization problem to a set of single-dimensional problems. This decomposition is not generally possible when one moves to non-Bayesian formulations and/or optimization criteria other than the expected discounted reward.

In this talk we will consider a frequentist version of the multi-armed-bandit model under the criterion of maximizing the expected average reward over an infinite horizon. We will show that the maximization problem is equivalent to the asymptotic minimization of an appropriately defined regret function. We will also prove the existence of index-based adaptive policies which ensure the minimum asymptotic rate

of increase of the regret, or equivalently the maximum rate of convergence of the average reward to that under complete information.

We will also consider two extensions of the basic model: (i) to the case of side constraints due to sampling costs and (ii) to the case of non stationarity of the reward structure.

The talk is based on joint research with Michael Katehakis (Rutgers University) and Odysseas Kanavetas (Sabanjee University).